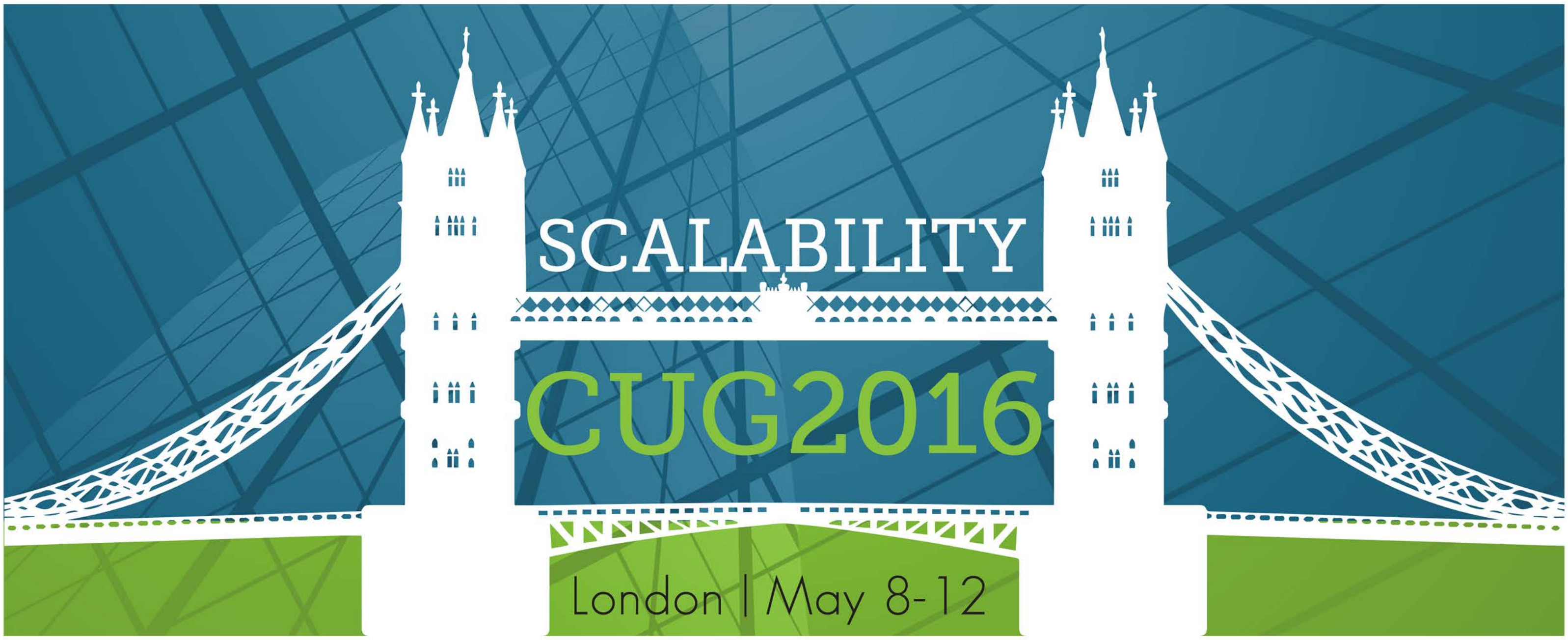


SCALABILITY

CUG2016

London | May 8-12

Welcome To CUG 2016

I am delighted to welcome all of you to London, for this 59th edition of the Cray User Group conference. Though ECMWF is a European organisation rather than a British one, the UK has been our host nation since the creation of the Centre in 1975, and as such we do consider it to be home to ECMWF.

As members of the Cray User Group for the past couple of years, we were delighted to be asked to host this edition, and choosing Scalability as a theme took us all of half a minute. Progress in numerical weather prediction is intimately connected with progress in supercomputing. Over the years, more computing power has enabled us to increase the skill and detail of our forecasts. This has brought huge value to society, not least through early warnings of severe weather. But as the forecasting system becomes more complex, with current computing architectures it will soon be impossible to issue forecasts within schedule and at a reasonable cost. The Scalability Programme that ECMWF has been leading for a few years is about making numerical weather prediction coding more efficient. We strongly believe that it is one of the biggest challenge our community has ever had to face. What makes it such a perfect fit for this gathering is that, like the Cray User Group, Scalability is about collaboration between users, computer scientists, software developers and hardware designers. It is about learning from each other, and about academia, operational centres and commercial sectors working together to ensure that technology and science can together achieve the best for society.

As members of the CUG, we have seen the benefits of engaging and sharing views with the diverse and broad range of members. The way we in NWP use the systems may be different from the way commercial users, say in the pharmaceutical industry use it, but sharing techniques and knowledge help us all. We've also enjoyed this special relationship it creates with Cray. Though clearly independent from the CUG, Cray does listen to such a strong group of users and this again benefits us all.

So a lot of good work to do during that week of CUG 2016, but don't forget that London is a very special city that you must take some time to discover or re-discover. I know the team have worked very hard to ensure that you can mix work and play, and I will just finish with thanking the CUG Board for all their hard work and their confidence in us!

Dr Florence Rabier
Director General, ECMWF

Invited Speakers

Petteri Taalas

Weather and Climate Services and need of High Performance Computing (HPC) Resources

We are entering a new era in technological innovation and in use and integration of different sources of information for the well-being of society and their ability to cope with multi-hazards through weather and climate services. New predictive tools that will detail weather conditions down to neighbourhood and street level, and provide early warnings a month ahead, and forecasts from rainfall to energy consumption will be some of the main outcome of the research activities in weather science over the next decade.

As weather and climate science advances, critical questions are arising such as about the possible sources of predictability on weekly, monthly and longer time-scales; seamless prediction; the development and application of new observing systems; the effective utilization of massively-parallel supercomputers; the communication, interpretation, and application of weather-related information; and the quantification of the societal impacts. The science is primed for a step forward informed by the realization that there can be predictive power on all space and time-scales arising from currently poorly-understood sources of potential predictability.

Globally the tendency of weather forecasting is moving towards impact-based direction. Besides forecasting the physical parameters, like temperature, wind and precipitation customers and general public are becoming more interested in the impacts of weather and climate phenomena. (Contiuned on page 15)

PhD in Meteorology, Helsinki Univ., physics department 1993
Management training: Helsinki Univ. Economics 1998 & 2004
Secretary-General of the World Meteorological Organization. Elected for a mandate from 2016 to 2019
Director-General of Finnish Meteorological Institute 2002-2005, 2007-2015

Invited Speakers

Florence Rabier

Florence Rabier first joined the European Centre for Medium-Range Weather Forecasts (ECMWF) as a consultant during her PhD in 1991, then as a scientist in data assimilation for 6 years in the 1990s. She came back from Météo-France in October 2013 to take up the position of Director of the newly formed Forecast Department. She is the Director General of ECMWF since January 2016.

Dr Rabier is also an internationally recognised expert in Numerical Weather Prediction, whose leadership has greatly contributed to delivering major operational changes at both ECMWF and Météo France. Dr Rabier is especially well known within the meteorological community for her key role in implementing a new data assimilation method (4D-Var) in 1997, which was a first worldwide. Her career so far has taken her back and forth between Météo France and ECMWF.
Dr Rabier has experience in both externally-funded and cross-departmental project management. She led an international experiment involving a major field campaign over Antarctica, in the context of the International Polar Year and THORPEX.
She has been awarded the title of "Chevalier de la Légion d'Honneur", one of the highest decorations awarded by the French honours system.

MONDAY MAY 9TH

08:30-10:00 Tutorial 1A

Cray Management System for XC Systems with SMW 8.0/CLE 6.0

Longley

New versions of CLE 8.0 and SMW 6.0 have been developed that include a new Cray Management System (CMS) for Cray XC systems. This new CMS includes system management tools and processes which separate software and configuration for the Cray XC systems, while at the same time preserving the system reliability and scalability upon which you depend. The new CMS includes a new common installation process for SMW and CLE, and more tightly integrates external login nodes (eLogin) as part of the Cray XC system. It includes the Image Management and Provisioning System (IMPS), the Configuration Management Framework (CMF), and Node Image Mapping Service (NIMS). Finally, it integrates with SUSE Linux Enterprise Server 12. The tutorial will cover an overview of the overall concepts of the new CMS, followed by examples of different system management activities.

08:30-10:00 Tutorial 1B

Knights Landing and Your Application: Getting Everything from the Hardware

Mallinson

Knights Landing, the 2nd generation Intel® Xeon Phi™ processor, utilizes many breakthrough technologies to combine breakthrough's in power performance with standard, portable, and familiar programming models. This presentation provides an in-depth tour of Knights Landing's features and focuses on how to ensure that your applications can get the most performance out of the hardware.

08:30-10:00 Tutorial 1C

CUG 2016 Cray XC Power Monitoring and Management Tutorial

Martin, Rush, Kappel, Williams

This half day (3 hour) tutorial will focus on the setup, usage and use cases for Cray XC power monitoring and management features. The tutorial will cover power and energy monitoring and control from three perspectives: site and system administrators working from the SMW command line, users who run jobs on the system, and third party software development partners integrating with Cray's RUR and CAPMC features.

13:00-14:30 Tutorial 2A

Cray Management System for XC Systems with SMW 8.0/CLE 6.0

Longley

New versions of CLE 8.0 and SMW 6.0 have been developed that include a new Cray Management System (CMS) for Cray XC systems. This new CMS includes system management tools and processes which separate software and configuration for the Cray XC systems, while at the same time preserving the system reliability and scalability upon which you depend. The new CMS includes a new common installation process for SMW and CLE, and more tightly integrates external login nodes (eLogin) as part of the Cray XC system. It includes the Image Management and Provisioning System (IMPS), the Configuration Management Framework (CMF), and Node Image Mapping Service (NIMS). Finally, it integrates with SUSE Linux Enterprise Server 12. The tutorial will cover an overview of the overall concepts of the new CMS, followed by examples of different system management activities.

13:00-14:30 Tutorial 2B

Getting the full potential of OpenMP on Many-core systems

Levesque, Poulsen

With the advent of the Knight’s series of Phi processors, code developers who employed only MPI in their applications will be challenged to achieve good performance. The traditional methods of employing loop level OpenMP are not suitable for larger legacy codes due to the risk of significant inefficiencies. Runtime overhead, NUMA effects, load imbalance are the principal issues facing the code developer. This tutorial will suggest a higher-level approach that has shown promise of circumventing these inefficiencies and achieving good performance on many-core systems. The approach takes a global view of the parallelization process, yet allows the programmer to exploit the design ideas behind OpenMP and complete the process incrementally. While this is technically more difficult and does require a good understanding of the language and OpenMP standard to avoid pitfalls inherent in the interaction between the two, the approach gives the programmer the ability to minimize the amount of synchronization, to control load balancing, to ensure a consistent NUMA layout of data and to assure that the thread data is local to the thread. We will show how these techniques have been used to refactor an ocean model successfully. We will repeat the techniques from the first part of the tutorial in this concrete context and eventually present the results that demonstrate the model is ready to run on Intel nodes of today as well as on Intel nodes that will emerge within a foreseeable future.

13:00-14:30 Tutorial 2C

eLogin Made Easy - An Introduction and Tutorial on the new Cray External Login Node

Keopp, Ebeling

The new eLogin product (external login node) is substantially different from its esLogin predecessor. Management is provided by the new OpenStack based Cray System Management Software. Images are prescriptively built with the same technology used to build CLE images for Cray compute and service nodes, and the Cray Programming Environment is kept separate from the operational image. This tutorial will provide information about the Cray eLogin product and features as well as best practices for the administration of eLogin nodes. This session will be ideal for administrators with hints and tips being provided for fully utilizing the features of eLogin and OpenStack.

16:45-18:00 Interactive 3A

Jobs I/O monitoring for Lustre at scale

Chesi, Hadri, Declerck, Trautmann, Hill, Petersen

In this session different Lustre users will talk about the monitoring tools available for Lustre Filesystems. The discussion will focus on the monitoring tools for user jobs I/O load in an attempt to evaluate their potentialities and limits concerning HPC systems scaling issues.

16:45-18:00 Interactive 3B

Containers for HPC

Canon, Jacobson, Alam

Container-based computing is an emerging model for developing and deploying applications and is now making inroads into the HPC community with the release of products like Shifter. The promise is large, but many challenges remain. What security concerns still need to be addressed? How do we train users to take advantage of this capability? What are best practices around creating and distributing images? How can we build up an ecosystem across the broader HPC community to promote reusability and increase productivity? This BOF will provide an opportunity to discuss these questions with experts from NERSC, CSCS, and Cray.

16:45-18:00 Interactive 3C

Open Discussion with CUG Board
Hancock

This session is designed as an open discussion with the CUG Board but there are a few high level topics that will also be on the agenda. The discussion will focus on corporation changes to achieve non-profit status (including bylaw changes), feedback on increasing CUG participation, and feedback on SIG structure and communication. An open floor question and answer period will follow these topics. Formal voting (on candidates and the bylaws) will open after this session, so any candidates or members with questions about the process are welcome to bring up those topics.

TUESDAY MAY 10TH

07:30-08:15 Interactive 4A

Sonexion Collaboration (Invitation Only)
Burke

07:30-08:15 Interactive 4B

Cray and HPC in the Cloud
Scott, Corbett, Kothari, Waite

Supercomputing customers have unique set of requirements and some fit better than others with standard cloud offerings. Many Cray customers have asked for cloud enablement and co-existence of their applications in the Cloud. Others provide de facto private and specialized cloud services to their own communities and customers. These sessions will drill down on use cases (current and anticipated), best practices, necessary infrastructure (SW/HW) developments and enablement for you to successfully deploy Cray for HPC in the cloud. Participants will be asked to participate in two consecutive BoF sessions. Session 1 will include a facilitator-led exercise to answer the determine and group participants' requirements for embracing cloud services in their operations. This session ends with and exercise. Session 2 collects the result of the exercise for focused discussion and prioritization of the findings. Results will be shared with the participants of this BoF. For continuity of discussion participants are asked to commit to both Sessions.

08:30-10:00 General Session 5

CUG Welcome
Hancock

08:30-10:00 General Session 5

The Strength of a Common Goal
Rabier

ECMWF is an intergovernmental organisation supported by 34 European States. It provides forecasts of global weather to 15 days ahead as well as monthly and seasonal forecasts. The National Meteorological Services of Member and Co-operating States use ECMWF's products for their own national duties, in particular to give early warning of potentially damaging severe weather. It operates a sophisticated prediction model of the global atmosphere and oceans known as the Integrated Forecasting System (IFS), running on some two and a half million lines of codes. Operational since 1994, the IFS is constantly updated to add new features to adapt to the latest HPC advances. ECMWF has been using supercomputers since 1977. It operates one of the largest supercomputer facilities of its type in Europe for meteorology worldwide and holds the world's largest archive of numerical weather prediction and observational data. Member and Co-operating States can access ECMWF's basic computing facilities, the meteorological archive, and temporary tape storage. Member States also have access to the supercomputers and permanent tape storage. Progress in numerical weather prediction is intimately connected with progress in supercomputing. Over the years, more computing power has enabled us to increase the skill and detail of our forecasts. This has brought huge value to society, not least through early warnings of severe weather. But as the forecasting system becomes more complex, with current computing architectures, it will soon be impossible to issue forecasts within schedule and at a reasonable cost. Supercomputer energy consumption at ECMWF would have to increase unviably, from about 4 megawatts today to perhaps 50 megawatts or more in ten years' time, if the more complex forecasting systems of the future were to be run on the current architecture. A new generation of computing systems with exascale capabilities promise much greater energy efficiency – but they will rely on parallel processing at levels to which current NWP codes are not adapted. Changes are needed throughout the entire NWP processing chain if we are to exploit these new opportunities for energy efficiency. ECMWF's Scalability Programme, launched in 2013, brings together meteorological modellers, computer scientists and hardware providers from around the world for a coordinated approach to hardware and software development. This ten-year programme encompasses the entire NWP processing chain, from processing and assimilating observational data to delivering forecasts to Member States. Projects will cover six main areas:

- Observational data processing
- Data assimilation
- Numerical methods
- Numerical data processing
- IFS code adaptation
- Computer architecture support

08:30-10:00 General Session 5

Numerical weather prediction ready to embrace Exascale!

Wedi

Presentation from the Head of Earth Modeling Section

08:30-10:00 General Session 5

Quo Vadis HPC?

Weger

Presentation from the Deputy Director of Computing

10:30-12:00 General Session 6

Cray Corporate Update

Ungaro

10:30-12:00 General Session 6

Cray Future Directions

Scott

10:30-12:00 General Session 6

Cray Products Update

Waite

13:00-14:30 Technical Session 7A

Performance on Trinity (a Cray XC40) with Acceptance-Applications and Benchmarks

Rajan, Wichmann, Nuss, Carrier, Olson, Anderson, Davis, Baker, Draeger, Domino, Agelastos

Abstract—Trinity is NNSA's first ASC Advanced Technology System (ATS) targeted to support the largest, most demanding nuclear weapon simulations. Trinity Phase-1 (the focus of this paper) has 9436 dual-socket Haswell nodes while Phase-2 will have close to 9500 KNL nodes. This paper documents the performance of applications and benchmarks used for Trinity acceptance. It discusses the early experiences of the Tri-Lab (LANL, SNL and LLNL) and Cray teams to meet the challenges for optimal performance on this new architecture by taking advantage of the large number of cores on the node, wider SIMD/vector units and the Cray Aries network. Application performance comparisons to our previous generation large Cray capability systems show excellent scalability. The overall architecture is facilitating easy migration of our production simulations to this 11 PFLOPS system, while improved work flow through the use of Burst-Buffer nodes is still under investigation.

13:00-14:30 Technical Session 7A

Improving I/O Performance of the Weather Research and Forecast (WRF) Model

Balle, Johnsen

As HPC resources continue to increase in size and availability, the complexity of numeric weather prediction models also rises. This increases demands on HPC I/O subsystems, which continue to cause bottlenecks in efficient production weather forecasting. In this paper we review the available I/O methodologies in the widely-used NCAR Weather Research and Forecasting (WRF) model. We focus on the newer PNETCDF_QUILT technique that uses asynchronous I/O (quilt) servers alongside Parallel NetCDF. This paper looks at a high-resolution nested WRF case and compares the performance of various I/O techniques. The PNETCDF_QUILT technique is then described in detail. The Cray implementation of MPI-IO and useful diagnostic settings are discussed. The focus is on the performance of WRF on the Cray XC40 platform, with both Sonexion 2000 and Cray DataWarp storage. The DataWarp results are some of the first available and will be interesting to a wide range of Cray users.

13:00-14:30 Technical Session 7A

Performance Evaluation of Apache Spark on Cray XC Systems

Chaimov, Allen, Ibrahim, Iancu, Canon, Srinivasan

We report our experiences in porting and tuning the Apache Spark data analytics framework on the Cray XC30 (Edison) and XC40 (Cori) systems, installed at NERSC. We find that design decisions made in the development of Spark are based on the assumption that Spark is constrained primarily by network latency, and that disk I/O is comparatively cheap. These assumptions are not valid on large HPC systems such as Edison or Cori, which feature advanced low-latency networks but have diskless compute nodes. The centralized storage system, such as Lustre, results in metadata access latency being a major bottleneck thus severely constraining scalability. We characterize this problem with benchmarks run on a system with both Lustre and local disks, and show how to mitigate high metadata access latency by using per-node loopback filesystems for temporary storage. With this technique, we reduce the communication (data shuffle) time and improve the application scalability from

O(100) to O(10,000) cores on Cori. For shuffle-intensive machine learning workloads, we show better performance than clusters with local disks.

13:00-14:30 Technical Session 7B

Finally, A Way to Measure Frontend I/O Performance

Zimmer, Vergara Larrea, Gupta

Identifying sources of variability in the Spider II file system on Titan is challenging because it spans multiple networks with layers of hardware performing various functions to fulfill the needs of the parallel file system. Several efforts have targeted file system monitoring but only focused on metric logging associated with the storage side of the file system. In this work, we enhance that view by designing and deploying a low-impact network congestion monitor designed especially for the IO routers that are deployed on service nodes within the Titan Cray XK7 Gemini network. To the best of our knowledge, this is is the first tool that provides a capability of live monitoring for performance bottlenecks at the IO router. Our studies show high correlation between IO router congestion and IO bandwidth. Ultimately, we plan on using this tool for IO hotspot identification within Titan and guided scheduling for large IO.

13:00-14:30 Technical Session 7B

A Classification of Parallel I/O Toward Demystifying HPC I/O Best Practices

Sisneros, Chadalavada

The process of optimizing parallel I/O can quite easily become daunting. By the nature of its implementation there are many highly sensitive, tunable parameters and a subtle change to any of these may have drastic or even completely counterintuitive results. There are many factors affecting performance: complex hardware configurations, significant yet unpredictable system loads, and system level implementations that perform tasks in unexpected ways. A final compounding issue is that an optimization is very likely specific to only a single application. The state of the art then is usually a fuzzy mixture of expertise and trial-and-error testing. In this work we introduce a characterization of application I/O based on aggregation which we define as a combination of job-level and filesystem-level. We will show how this characterization may be used to analyze parallel I/O performance to not only validate I/O best practices but also communicate benefits in a user centric way.

13:00-14:30 Technical Session 7B

Collective I/O Optimizations for Adaptive Mesh Refinement Data Writes on Lustre File System

Devendran, Byna, Dong, Van Straalen, Johansen, Keen, Samatova

Adaptive mesh refinement (AMR) applications refine small regions of a physical space. As a result, when AMR data has to be stored in a file, writing data involves storing a large number of small blocks of data. Chombo is an AMR software library for solving partial differential equations over block-structured grids, and is used in large-scale climate and fluid dynamics simulations. Chombo's current implementation for writing data on an AMR hierarchy uses several independent write operations, causing low I/O performance. In this paper, we investigate collective I/O optimizations for Chombo's write function. We introduce Aggregated Collective Buffering (ACB) to reduce the number of small writes. We demonstrate that our approach outperforms the current implementation by 2X to 9.1X and the MPI-IO collective buffering by 1.5X to 3.4X on the Edison and Cori platforms at NERSC using the Chombo-IO benchmark.

Using the Darshan I/O characterization tool, we show that ACB makes larger contiguous writes than collective buffering at the POSIX level, and this difference gives ACB a significant performance benefit over collective buffering.

13:00-14:30 Technical Session 7C

Unified Workload Management for the Cray XC30/40 System with Univa Grid Engine

Gruber, Ferstl

Workload management (WLM) software provides batch queues and scheduling intelligence for efficient management of Cray systems. The widely used Univa Grid Engine (UGE) WLM software is available and in commercial production use for several years now for Cray XC30/40 systems. UGE allows to integrate Cray system seamlessly with other computational resources of an organization to form one, unified WLM domain. This paper describes the general structure and features of UGE, how those components map onto a Cray system, how UGE is integrated with the Cray infrastructure, and what the user will see for job structure and features in using Univa Grid Engine. Special features of UGE are also highlighted.

13:00-14:30 Technical Session 7C

Driving More Efficient Workload Management on Cray Systems with PBS Professional

Russell, Alhabashneh

The year 2015 continued to increase the adoption of key HPC technologies, from data analytics solutions to power-efficient scheduling. The HPC user landscape is changing and it is now critical for workload management vendors to provide not only foundational scheduling functionality but also the adjacent capabilities that truly optimize system performance. In this presentation, Altair will provide a look at key advances in PBS Professional for improved performance on Cray systems. Topics include new Cray-specific features like Suspend/Resume, Xeon Phi support, Power-aware Scheduling, and DataWarp.

13:00-14:30 Technical Session 7C

Broadening Moab for Expanding User Needs

Brown

Adaptive Computing's Moab HPC scheduler, Nitro HTC scheduler, and the open-source TORQUE RM are broadening their reach to handle increasingly diverse users and their needs. This presentation will discuss the following examples. - Moab's integration with Cray's speedy SSD-based DataWarp for the Trinity system - Moab and TORQUE support for Intel's new Xeon Phi "KNL" with its "near" memory and multiple NUMA node configurations - Nitro's fast execution of small HTC jobs under any job scheduler with performance measurements - Moab Viewpoint's web browser-based User Portal for technical and non-technical users, with additional support for Nitro and interactive Remote Visualization jobs - Moab Elastic Compute's automatic management of virtual HPC clusters on physical HPC resources supporting full isolation between medical research organizations, projects, and researchers using private VMs for compliance with privacy regulations.

15:00-17:00 Technical Session 8A

Early Experiences Writing Performance Portable OpenMP 4 Codes

Vergara Larrea, Joubert, Lopez, Hernandez

At least two major architectural trends are leading the way to Exascale: accelerator-based (e.g., Summit and Sierra), and self-hosted compute nodes (e.g., Aurora). Today, the ability to produce performance portable code is crucial to take full advantage of these different architectures. Directive-based programming APIs (e.g., OpenMP, OpenACC) have helped in this regard, and recently, OpenMP added an accelerator programming model, in addition to its shared memory programming model support. However, as of today, little is understood about how efficiently the accelerator programming model can be mapped onto different architectures, including self-hosted and traditional shared memory systems, and whether it can be used to generate performance portable code across architectures. In this paper, we parallelize a representative computational kernel using the two different OpenMP 4 styles (shared memory and accelerator models), and compare their performance on multiple architectures including OLCF's Titan supercomputer.

15:00-17:00 Technical Session 8A

A Knowledge Reasoning and Hypothesis Generation Framework using Urika-XA and Urika-GD

Sukumar, Roberts, Graves, Rogers

Finding actionable insights from data has always been difficult. As the scale and forms of data increase tremendously, the task of finding value becomes even more challenging. Data scientists at Oak Ridge National Laboratory are leveraging unique leadership infrastructure (e.g. Urika-XA and Urika-GD appliances) to develop scalable algorithms for semantic, logical and statistical reasoning with unstructured Big Data. We present the deployment of such a framework called ORiGAMI (Oak Ridge Graph Analytics for Medical Innovations) on the National Library of Medicine's SEMANTIC Medline (archive of medical knowledge since 1994). Medline contains over 70 million knowledge nuggets published in 23.5 million papers in medical literature with thousands more added daily. ORiGAMI is available as an open-science medical hypothesis generation tool - both as a web-service and an application programming interface (API) at http://hypothesis.ornl.gov

15:00-17:00 Technical Session 8A

Code Porting to Cray XC-40 Lesson Learned

McClean, Gautam

We present a case study of porting seismic applications from the Beowulf cluster using Ethernet network to the Cray XC40 cluster. The applications in question are Tilted Transverse Anisotropic Reverse Time Migration (TTI RTM), Kirchhoff Depth Migration (KDMIG) and Wave equation migration (WEM). The primary obstacle in this port was that TTI RTM and WEM use local scratch disk heavily and imaging is performed one shot per node. The Cray nodes do not have local scratch disks. The primary obstacle in KDMIG was its heavy IO usage from permanent disk due to the constant reading of Travel Time Maps. We briefly explain how these algorithms were refactored so as to not be primarily dependent on scratch disk and also to fully utilize the better networking in the Cray XC40. In the case of KDMIG, we explain how its IO load was reduced via a memory pool concept.

15:00-17:00 Technical Session 8A

Trinity: Architecture and Early Experience

Hemmert, Vigil, Lujan, Hoekstra, Grunau, Morton, Nam, Peltz, Jr., Torrez, Wright, Dawson

The Trinity supercomputer is the first in a series of Advanced Technology Systems (ATS) that will be procured by the DOE's Advanced Simulation and Computing program over the next decade. The ATS systems serve the dual role of meeting immediate mission needs and helping to prepare for future system designs. Trinity meets this goal though a two-phase delivery. Phase 1 consists of Xeon-based compute nodes, while phase 2 adds Xeon Phi-based nodes. Phase 1 was delivered and stood-up during the end of 2015. This paper describes early experiences and performance evaluations from Trinity phase 1. Early results are promising; phase 1 has met, and often exceeded, the required performance set forth in acceptance criteria, in the areas of capability improvement (CI), system sustained performance (SSP) and filesystem performance, among others. This paper describes the Trinity architecture, performance evaluations, early experience with system management and application readiness efforts.

15:00-17:00 Technical Session 8B

Improving User Notification on Frequently Changing HPC Environments

Fuson, Renaud, Wynne III

Today's HPC centers' user environments can be very complex. Centers often contain multiple large complicated computational systems each with their own user environment. Changes to a system's environment can be very impactful; however, a center's user environment is, in one-way or another, frequently changing. Because of this, it is vital for centers to notify users of change. For users, untracked changes can be costly, resulting in unnecessary debug time as well as wasting valuable compute allocations and research time. Communicating frequent change to diverse user communities is a common and ongoing task for HPC centers. This paper will cover the OLCF's current processes and methods used to communicate change to users of the center's large Cray systems and supporting resources. The paper will share lessons learned and goals as well as practices, tools, and methods used to continually improve and reach members of the OLCF user community.

15:00-17:00 Technical Session 8B

Technical Publications New Portal

Sanchez

Over the last year and a half, the Cray Publications department has gone through revolutionary changes. The new Cray Portal (not to be confused with CrayPort the customer service portal) is a tool based on the results of these changes. There are Cray internal benefits but more importantly are the user benefits due to the new standard. The portal is innovative for technical publications and a sought after example of new technology. While the portal is relatively simple to use, it is a rare, albeit excellent model of what users can do within the same standards. Key benefits of this presentation will be for the attendees to understand how the documentation portal brings opportunities to customize, rate, and respond to content as well as see the future of content delivery in a responsive design at Cray.

15:00-17:00 Technical Session 8B

Slurm Overview and Road Map

Jenson

Slurm is an open source workload manager used on five of the world's top 10 most powerful computers and provides a rich set of features including topology aware optimized resource allocation, the ability to expand and shrink jobs on demand, failure management support for applications, hierarchical bank accounts with fair-share job prioritization, job profiling, and a multitude of plugins for easy customization. This presentation will provide a brief overview of Slurm and review enhancements made to Slurm 15.08 and 16.05 as well as provide a road map for Slurm 17.02.

15:00-17:00 Technical Session 8B

Interactive Visualization of Scheduler Usage Data

Gall

This paper describes the use of contemporary web client-based interactive visualization software to explore HPC job scheduler usage data for a large system. Particularly, we offer a visualization web application to NOAA users that enables them to compare their experiences with their earlier experiences and those of other users. The application draws from a 30 day history of job data aggregated hourly by user, machine, QoS, and other factors. This transparency enables users to draw more informed conclusions about the behavior of the system. The technologies used are dc.js, d3.js, crossfilter, and bootstrap. This application differs from most visualizations of job data in that we eschewed the absolute time domain to focus on developing relevant charts in other domains, like relative time, job size, priority boost, and queue wait time. This enables us to more easily see repetitive trends in the data like diurnal and weekly cycles of job submissions.

15:00-17:00 Technical Session 8C

Crossing the Rhine - Moving to CLE 6.0 System Management

Butler, Declerck

With the release of Cray Linux Environment 6.0, Cray has introduced a new paradigm for CLE system configuration and management. This major shift requires significant changes in formatting and practices on the System Management Workstation (SMW). Although Cray has committed to delivering migration tools for legacy systems, they will not be available until CLE 6.0 UP02, scheduled for July 2016 release. In the third quarter of 2016, NERSC will be taking delivery of the second phase of its Cori system, with Intel KNL processors. KNL requires CLE 6.0. In order to support phase 2, Cori will have to be upgraded to CLE 6.0 - the hard way. This paper will chronicle that effort.

15:00-17:00 Technical Session 8C

Making the jump to Light Speed with Cray's DataWarp - An Administrator's Perspective

Declerck, Paul

Cori, the first phase of NERSC's next generation supercomputer, has 144 DataWarp nodes available to it's users. Cray's DataWarp technology provides an intermediate storage capability, sitting between on node memory and the parallel file system. It utilizes Cray's DVS to provide access from compute nodes on a request basis. In order for this to work, the workload manager interacts with Cray's DataWarp API's to create the requested file system and make it available on the nodes requested by the job (or jobs). Some of the tools needed by an administrator are therefore included in the workload manager, SLURM at our site, and other information requires use of tools and commands provided by Cray. It is important to know what information is available, where to find it, and how to get it.

15:00-17:00 Technical Session 8C

The NERSC Data Collect Environment

Whitney, Bautista, Davis

As computational facilities prepare for Exascale computing, there is a wider range of data that can be collected and analyzed but existing infrastructures have not scaled to the magnitude of the data. Further, as systems grow, there is wider impact of their environmental footprint and data analysis should include answers to power consumption, a correlation to jobs processed and power efficiency as well as how jobs can be scheduled to leverage this data. At NERSC, we have created a new data collection methodology for the Cray system that goes beyond the system and extends into the computational center. This robust and scalable system can help us manage the center, the Cray and ultimately help us learn how to scale our system and workload to the Exascale realm.

15:00-17:00 Technical Session 8C

Scaling Security in a Complex World

Palm

Cray systems are becoming increasingly complex while security awareness has intensified. This presentation will address the changes in Cray's security processes that have resulted from this increased focus as well as our new installation and update procedures.

17:15-18:15 Interactive 9A

Systems Support SIG Meeting

Hill

17:15-18:15 Interactive 9B

Best Practices for Managing HPC User Documentation and Communication

Fuson, Barker, Richard, Indiviglio

HPC centers provide large, complex, state-of-the-art computational and data resources to large user communities that span diverse science domains and contain members with varied experience levels. Effectively using these resources can pose a challenge to users, especially considering that each center often has site-specific configurations and procedures. To ensure the wide range of users who have been granted limited time on a center's HPC resources have the ability to use the resources effectively, it is imperative that centers provide accurate and extensive documentation as well as communicate changes that will impact a user's workflow. This can pose a challenge to HPC centers because managing hundreds of pages of documentation covering a wide range of topics and audiences as well as communicating to large numbers of users can be substantial efforts. Finding the correct expert to create documentation, ensuring documentation remains up-to-date, creating web navigation, and finding ways to reach busy users are just some of the tasks that complicate the effort. While many CUG member sites share common communication and documentation goals and a similar Cray computing environment on many large-scale systems, collaboration between centers on their efforts is often not as common. The goal of this BOF is provide an open forum for those involved in user communication and documentation to discuss tools and methods used as well as share lessons learned and best practices. The session will also be used to discuss the possibility for future collaboration efforts among both the centers and Cray.

17:15-18:15 Interactive 9C

GPU accelerated Cray XC systems: Where HPC meets Big Data

Messmer, Lindahl, Alam

We discuss accelerated compute, data analysis and visualization capabilities of NVIDIA Tesla GPUs within the scalable and adaptable Cray XC series supercomputers. Historically, HPC and Big Data had their distinct challenges, algorithms and computing hardware. Today, the amount of data produced by HPC applications turns their analysis into a Big Data challenge. On the other hand, the computational complexity of modern data analysis requires compute and messaging performance only found in HPC systems. The ideal supercomputer therefore unites compute, analysis and visualization capabilities. Presenters from NVIDIA, Cray and HPC sites will showcase features of the XC series that go beyond accelerated computing and demonstrate how heterogeneous systems are the ideal platform for converging high-end computing and data analysis. We will cover new features of NVIDIA drivers and libraries, allowing to leverage the GPU's graphics capabilities, Cray's support for container technologies like shifter, and discuss an integrated and yet decoupled programming and execution environment as a highly performance and yet flexible platform for a wide range of traditional HPC but also emerging analytics and data science applications and workflows.

WEDNESDAY MAY 11TH

07:30-08:15 Interactive 10A
Addressing the challenges of "systems monitoring" data flows
Showerman, Brandt, Gentile

As Cray systems have evolved in both scale and complexity, the volume of quality systems data has grown to levels that are challenging to process and store. This BOF is an opportunity to discuss some of the use cases for high resolution power, interconnect, compute, and storage subsystem data. We hope to be able to gain insights into the requirements sites have for future systems deployments, and how these data need to be integrated and processed. There will be presentations of known problems that cannot be addressed with the current infrastructure design, as well as directions Cray could go to meet the needs of sites.

07:30-08:15 Interactive 10B
Technical Documentation and Users
Sanchez

An open discussion about the direction of documentation for users as part of the continuing effort to improve the user experience and better understand the needs outside of Cray. This is an opportunity to contribute to the direction of content within technical publications as well as identify immediate needs.

08:30-10:00 General Session 11
CUG Business
Hancock

CUG President's Report
CUG Treasurer's Report
SIG Reports CUG Elections

08:30-10:00 General Session 11
Weather and Climate Services and need of High Performance Computing (HPC) Resources
Taalas

We are entering a new era in technological innovation and in use and integration of different sources of information for the well-being of society and their ability to cope with multi-hazards through weather and climate services. New predictive tools that will detail weather conditions down to neighbourhood and street level, and provide early warnings a month ahead, and forecasts from rainfall to energy consumption will be some of the main outcome of the research activities in weather science over the next decade. As weather and climate science advances, critical questions are arising such as about the possible sources of predictability on weekly, monthly and longer time-scales; seamless prediction; the development and application of new observing systems; the effective utilization of massively-parallel supercomputers; the communication, interpretation, and application of weather-related information; and the quantification of the societal impacts. The science is primed for a step forward informed by the realization that there can be predictive power on all space and time-scales arising from currently poorly-understood sources of potential predictability. Globally the tendency of weather forecasting is moving towards impact-based direction. Besides forecasting the physical parameters, like temperature, wind and precipitation customers and general public are becoming more interested in the impacts of weather and climate phenomena. These are for example traffic disturbances, impacts on energy availability and demand, impacts on agriculture and tourism. In the case of climate services the customers are e.g. investors, industry, insurance sector, construction companies, consumer businesses and agriculture. Besides the impact-based forecasting there is a tendency to move towards multi-hazard forecasting. As an example the Japanese earthquake led to a severe tsunami, which caused the Fukushima nuclear accident with dispersion of radioactive release to both atmosphere and ocean. The further evolution of HPCs is having an impact on weather and climate service providers globally. Most of the national services cannot afford the largest available computers for running high-resolution NWP and climate models with advanced physics. A key question is the possibility to get access to products provided by using the best models run at the largest HPCs. At the moment several countries and the European Commission are moving towards open data policies with free access to NWP model data. This allows also countries without own HPC resources and modelling skills to get access to state of the art products. On the other hand the growing interest of private sector in NWP modelling may have an opposite impact. The global strength of meteorological community so far has been relatively free exchange of know-how and the ability to collaborate across national borders, which is a key factor behind the success of the ECMWF. At the moment the national weather services are considering how to ensure access to better HPC resources. For example the five Nordic and three Baltic countries have agreed to seek for a joint HPC solution to enhance the value for money of national HPC investment resources. It is also discussed whether the countries should in the future buy own HPCs or rather get access to resource owned by e.g. private sector.

10:30-12:00 General Session 12

Accelerating Science with the NERSC Burst Buffer Early User Program

Bhimji, Bard, Romanus AbdelBaky, Paul

NVRAM-based Burst Buffers are an important part of the emerging HPC storage landscape. The National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory recently installed one of the first Burst Buffer systems as part of its new Cori supercomputer, collaborating with Cray on the development of the DataWarp software. NERSC has a diverse user base comprised of over 6500 users in 750 different projects spanning a wide variety of scientific applications, including climate modeling, combustion, fusion, astrophysics, computational biology, and many more. The potential applications of the Burst Buffer at NERSC are therefore also considerable and diverse. We describe here the Burst Buffer Early User Program at NERSC, which selected a number of research projects to gain early access to the Burst Buffer and exercise its different capabilities to enable new scientific advancements. We present details of the program, in-depth performance results and lessons-learnt from highlighted projects.

10:30-12:00 General Session 12

Is Cloud A Passing Fancy?

Hazra

Enterprise businesses are rapidly moving to the cloud, transforming from bricks and mortar into cloud based services. They increasingly rely on high volume data collection enabling complex analysis and modelling as fundamental business capabilities. Is this shift to Cloud good or bad for supercomputing? Join Raj as he discusses what cloud unique requirements mean for industry focus, and how investments will change, accelerating open source software and innovation models.

13:00-13:45 General Session 13

1 on 100 or More...

Ungaro

Open discussion with Cray President and CEO. No other Cray employees or Cray partners are permitted during this session.

14:00-15:00 Technical Session 14A

Estimating the Performance Impact of the MCDRAM on KNL Using Dual-Socket Ivy Bridge nodes on Cray XC30

Zhao

NERSC is preparing for its next petascale system, Cori, a Cray XC system which will have over 9300 KNL manycore processor nodes with 72 cores per node, 512 bit vector units, and the on-package high bandwidth memory (HBM or MCDRAM). To take advantage of these new features, NERSC has developed the optimization strategies that focus on the MPI+OpenMP program model, vectorization, and the efficient use of the HBM. While the optimization on MPI+OpenMP and vectorization can be performed on today's multi-core architectures, the optimization on the efficient use of HBM is difficult to perform on today's architecture where no HBM is available. In this paper, we will present our HBM performance analysis on the VASP code, a widely used material science code, using the Intel's development tools, Memkind and AutoHBW, using the dual-socket Ivy Bridge processor node on Edison, a Cray XC30, as the proxy to the HBM on KNL.

14:00-15:00 Technical Session 14A

Cray Performance Tools Enhancements for Next Generation Systems

Poxon

The Cray performance tools provide a complete solution from instrumentation, measurement, analysis and visualization of data. The focus of the tools is on whole program analysis, providing insight into performance bottlenecks within programs that use many computing resources across many nodes. With two complimentary interfaces: one for first-time users that provides a program profile at the end of program execution, and one for advanced users that provides in-depth performance investigation and tuning assistance, the tools enable users to quickly identify areas in their programs that most heavily impact performance or energy consumption. Recent development activity targets the new Intel KNL many-core processors, more assistance with adding OpenMP to MPI programs, improved tool usability, and enhanced application power and energy monitoring feedback. New CrayPat, Reveal and Cray Apprentice2 functionality is presented that will offer additional insight into application performance on next generation Cray systems.

14:00-15:00 Technical Session 14B

Architecture and Design of Cray DataWarp

Henseler, Landsteiner, Petesch, Wright

This paper describes the architecture, design, use and performance of Cray DataWarp, an infrastructure that uses direct attached solid state disk (SSD) storage to provide more cost effective bandwidth than an external parallel file system (PFS), allowing DataWarp to be provisioned for bandwidth and the PFS to be provisioned for capacity and resiliency. Placing this storage between the application and the PFS allows application I/O to be decoupled from (and in some cases eliminating) PFS I/O. This reduces the time required for the application to do I/O, while also increasing the overlap of computation with PFS I/O and typically reducing application elapsed time. DataWarp allocates and configures SSD backed storage for jobs and users on demand, providing many of the benefits of both software defined storage and storage virtualization.

14:00-15:00 Technical Session 14B

Exascale HPC Storage – A possibility or a pipe dream?

Petersen

The advances in flash and NV-RAM technologies promise exascale level throughput, however, the building and implementing of full solutions continues to be expensive. Noting HDDs are increasing in capacity and speed, are these new drives good enough to fulfill these essential areas? Many suggest a combination but that suggests software capable of handling multi-level storage transparently; such software does not actually exist in the HPC world today. What's the use of terabyte per second scratch filesystems when the second tier works at megabyte per second rates? There are more and more requests for enterprise reliability and supportability, end to end system monitoring and predictive failure, data management functionality and power efficiency. Pulling together all these requisites into a single solution is a tall order indeed. This talk discusses these areas in relation to the Seagate ClusterStor solutions – where we are today and where it might go in the future.

14:00-15:00 Technical Session 14C

ACES and Cray Collaborate on Advanced Power Management for Trinity

Laros III, Pedretti, Olivier, Grant, Levenhagen, Debonis, Pakin, Falde, Martin, Kappel

The motivation for power and energy measurement and control capabilities for High Performance Computing (HPC) systems is now well accepted by the community. While technology providers have begun to deliver some features in this area, interfaces to expose these features are vendor specific. The need for a standard interface, now and in the future is clear. To address this need, DOE funded an effort to produce a Power API specification for HPC systems with the goal of contributing this API to the community as a proposed standard for power measurement and control. In addition to the open publication of this standard an APM NRE project has been initiated with Cray Inc. with the intention of advancing capabilities in this area. We will detail the collaboration established between ACES and Cray and the portions of the Power API that have been selected for the first production implementation of the standard.

14:00-15:00 Technical Session 14C

Cray XC Power Monitoring and Control for Knights Landing (KNL)

Martin, Rush, Kappel, Sandstedt, Williams

This paper details the Cray XC40 power monitoring and control capabilities for Intel Knights Landing (KNL) based systems. The

Cray XC40 hardware blade design for Intel KNL processors is the first in the XC family to incorporate enhancements directly related to power monitoring feedback driven by customers and the HPC community. This paper focuses on power monitoring and control directly related to Cray blades with Intel KNL processors and the interfaces available to users, system administrators, and workload managers to access power management features.

15:30-17:00 Technical Session 15A

Lonestar 5: Customizing the Cray XC40 Software Environment

Proctor, Gignac, McLay, Liu, James, Minyard, Stanzione

Lonestar 5, a 30,000 core, 1.2 petaflop Cray XC40, entered production at the Texas Advanced Computing Center (TACC) on January 12, 2016. Customized to meet the needs of TACC's diverse computational research community, Lonestar 5 provides each user a choice between two alternative, independent configurations. Each is robust, mature, and proven: Lonestar 5 hosts both the environment delivered by Cray, and a second customized environment that mirrors Stampede, Lonestar 4, and other TACC clusters. This paper describes our experiences preparing Lonestar 5 for production and transitioning our users from existing resources. It focuses on unique features of the system, especially customizations related to administration (e.g. hierarchical software stack; secure virtual login nodes) and the user environment (e.g. consistent, full Linux environment across batch, interactive compute, and login sessions). We motivate our choices by highlighting some of the particular needs of our research community.

15:30-17:00 Technical Session 15A

The Cray Programming Environment: Current Status and Future Directions

DeRose

In this talk I will present the recent activities, roadmap, and future directions of the Cray Programming Environment, which are being developed and deployed on Cray Clusters and Cray Supercomputers for scalable performance with high programmability. The presentation will discuss Cray's programming environment new functionality to help porting and hybridizing applications to support systems with Intel KNL processors. This new functionality includes compiler directives to access high bandwidth memory, new features in the scoping tool Reveal to assist in parallelization of applications, and the Cray Comparative Debugger, which was designed and developed to help identify porting issues. In addition, I will present the recent activities in the Cray Scientific Libraries, and the Cray Message Passing Toolkit, and will discuss the Cray Programming Environment strategy for accelerated computing with GPUs, as well as the Cray Compiling Environment standard compliance plans for C++14, OpenMP 4.5, and OpenACC.

15:30-17:00 Technical Session 15A

Making Scientific Software Installation Reproducible On Cray Systems Using EasyBuild

Forai, Hoste, Peretti Pezzi, Bode

Cray provides a tuned and supported OS and programming environment (PE), including compilers and libraries integrated with the modules system. While the Cray PE is updated frequently, tools and libraries not in it quickly become outdated. In addition, the amount of tools, libraries and scientific applications that HPC user support teams are expected to provide support for is increasing significantly. The uniformity of the software environment across Cray sites makes it an attractive target for to share this ubiquitous burden, and to collaborate on a common solution. EasyBuild is an open-source, community-driven framework to automatically and reproducibly install (scientific) software on HPC systems. This paper presents how EasyBuild has been integrated with the Cray PE, in order to leverage the provided optimized components and support an easy way to deploy tools, libraries and scientific applications on Cray systems. We will discuss the changes that needed to be made to EasyBuild to achieve this, and outline the use case of providing the Python distribution and accompanying Python packages on top of the Cray PE, to obtain a fully featured `batteries included' optimized Python installation that integrates with the Cray-provided software stack. In addition, we will outline how EasyBuild was deployed at the Swiss National Supercomputing Centre CSCS and how it is leveraged to obtain a consistent software stack and installation workflow across multiple Cray (XC, CS-Storm) and non-Cray HPC systems.

15:30-17:00 Technical Session 15B

The Evolution of Lustre Networking at Cray

Horn

Lustre Network (LNet) routers with more than one InfiniBand Host Channel Adapter (HCA) have been in use at Cray for some time. This type of LNet router configuration is necessary on Cray Supercomputers in order to extract maximum performance out of a single LNet router node. This paper provides a look at the state of the art in this dual-HCA router configuration. Topics include avoiding ARP flux with proper subnet configuration, flat vs. fine-grained routing, and configuration emplacement. We'll also provide a look at how LNet will provide compatibility with InfiniBand HCAs requiring the latest mlx5 drivers, and what is needed to support a mix of mlx4 and mlx5 on the same fabric.

15:30-17:00 Technical Session 15B

Managing your Digital Data Explosion

Starr, Kinnin

Our society is currently undergoing an explosion in digital data. It is predicted that our digital universe will double every two years to reach more than 44 zettabytes (ZB) by 2020. The volume of data created each day has increased immensely and will continue to grow exponentially over time. This trend in data growth makes it clear that the data storage problems we struggle with today will soon seem very minor. So, where are we going to put all of this data, and how are we going to manage it? Short-term fixes are often worsening long-term storage challenges associated with performance, capital cost, operating cost and floor space. In today's talk, we will examine several unique storage medium options including tape, disk and cloud. Learn how you can affordably, efficiently and reliably manage your data, and scale to meet the needs of your explosive content growth.

15:30-17:00 Technical Session 15B

Extreme Scale Storage & IO

Barton

New technologies such as 3D Xpoint and integrated high performance fabrics are set to revolutionize the storage landscape as we reach towards Exascale computing. Unfortunately the latencies inherent in today's storage software mask the benefits of these technologies and the horizontal scaling. This talk will describe the work currently underway in the DOE funded Extreme Scale Storage & IO project to prototype a storage stack capable of exploiting these new technologies to the full and designed to overcome the extreme scaling and resilience challenges presented by Exascale Computing.

15:30-17:00 Technical Session 15C

Analysis of Gemini Interconnect Recovery Mechanisms: Methods and Observations

Jha, Formicola, Di Martino, Kalbarczyk, Kramer, Iyer

This paper presents methodology and tools to understand and characterize the recovery mechanisms of the Gemini interconnect system from raw system logs. The tools can assess the impact of these recovery mechanisms on the system and user workloads. The methodology is based on the topology-aware state-machine based clustering algorithm to coalesce the Gemini-related events (i.e., errors, failure and recovery events) into groups. The presented methodology has been used to analyze more than two years of logs from Blue Waters, the 13.1-petaflop Cray hybrid supercomputer at the University of Illinois - National Center for Supercomputing Applications (NCSA).

15:30-17:00 Technical Session 15C

SLURM. Our way. A tale of two XCs transitioning to SLURM.

Jacobsen, Botts, He

NERSC recently transitioned its batch system and workload manager from an ALPS based solution to SLURM running “natively” on our Cray XC systems. The driving motivation for making this change is to gain access to features NERSC has long implemented in alternate forms, such as a capacity for running large numbers of serial tasks, and gaining tight user-interface integration with new features of our systems, such BurstBuffers, Shifter, and VTune, while still retaining access to a flexible batch system that delivers high utilization of our systems. While we have derived successes in all these areas, perhaps the largest unexpected impact has been the change in how our staff interacts with the system. Using SLURM as the native WLM has blurred the line between system management and operation. This has been greatly beneficial in the impact our staff have on system configuration and deployment of new features: a platform for innovation.

15:30-17:00 Technical Session 15C

Early experiences configuring a Cray CS Storm for Mission Critical Workloads

Klein, Induni

MeteoSwiss is transitioning from a traditional Cray XE6 system to a very dense GPU configuration of the Cray CS Storm. This paper will discuss some of the system design choices and configuration decisions that has gone into the new setup in order to operate the mission critical workloads of weather forecasting. This paper will share some of the modifications that have been made to enhance things such as CPU/GPU/HCA affinity in the job scheduler, monitoring systems that have been set up to examine performance fluctuations, and also discuss the design of the failover system. This paper will also share some challenges found with the CS Storm management software, as well as the current support situation for this product line.

17:15-18:15 Interactive 16A

Programming Environments, Applications and Documentation SIG Meeting

Robinson

17:15-18:15 Interactive 16B

Evolving parallel file systems in response to the changing storage and memory landscape

Spitz

Burst buffers and storage-memory hierarchies are disruptive technologies to parallel file systems (PFS), but there isn't consensus among members of the HPC community on how PFSs should adapt to include their use, if at all. There also isn't consensus on how HPC users should ultimately use and manage their data with these emerging technologies. In this BoF we will discuss how HPC and technical computing users want to interact with burst buffers or other storage-memory hierarchies and how their PFS should adapt. What do they expect? Will they want to continue to use POSIX-like semantics for access like with MPI-I/O or HDF5 containers? What do users expect for legacy codes? Generally, what do users, application developers, and systems engineers require? Will they accept exotic solutions or must a de facto industry standard emerge?

17:15-18:15 Interactive 16C

Security Issues on Cray Systems

Palm

Cray sites and systems have a variety of functions, features and security requirements. This BoF is an opportunity for sites to contribute to the future plans, focus and direction of the Cray Security Team. Discussions of past security events will be used as examples.

THURSDAY MAY 12TH

07:30-08:15 Interactive 17A

PBS Professional: Welcome to the Open Source Community

Nitzberg

Altair will be releasing PBS Pro under an Open Source license in mid-2016. Join us to learn more and meet with others interested in open source PBS Pro. After a (very short) introduction, this BOF will be open discussion and feedback. In particular, the goal of this BOF is two-fold: engaging those interested in participating in the open source PBS Pro effort, and providing a forum for honest advice and feedback on our open source plans. Background: One reason that no clear winner has emerged from the dozens of different choices of HPC job scheduling is that there is a dichotomy in the HPC world: the public sector, with early adopters and risk takers who strongly favor open source software; and the private sector, with later adopters and the risk-averse who strongly favor commercial software. Because the community is split, innovations do not easily flow from one sector to the other, and this is a huge missed opportunity. This BOF will provide a forum to discuss the opportunity that Altair's dual-licensing approach to PBS Professional® software & OpenHPC have to unite the whole HPC community and greatly accelerate the state of the art (and the state of actual production operations) for scheduling.

07:30-08:15 Interactive 17B

Cray and HPC in the Cloud: Discussion and Conclusion

Scott, Kothari, Corbett, Waite

The BoF session follows up on and concludes a previous BoF session dedicated to Cray and HPC in the Cloud. (For continuity of discussion, participants in this BoF are asked to commit to the previous BoF.) This BoF collects the result of the previous exercise for focused discussion. Cray's roadmap and related development efforts for HPC in the Cloud will be presented and positioned against those findings for affinity, ability ranking and prioritization. Session will conclude with a "next steps" action plan.

08:30-10:00 Technical Session 18A

Opportunities for container environments on Cray XC30 with GPU devices

Benedicic, Gila, Alam

Thanks to the significant popularity gained lately by Docker, HPC community have recently started exploring container technology and the potential benefits its use would bring to the users of supercomputing systems like the Cray XC series. In this paper, we explore feasibility of diverse, nontraditional data and computing oriented use cases with practically no overhead thus achieving native execution performance. Working in close collaboration with NERSC and an engineering team at Nvidia, CSCS is working on extending the Shifter framework in order to enable GPU access to containers at scale. We also briefly discuss the implications of using containers within a shared HPC system from the security point of view to provide service that does not compromise the stability of the system or the privacy of the use. Furthermore, we describe several valuable lessons learned through our analysis and share the challenges we encountered.

08:30-10:00 Technical Session 18A

Shifter: Containers for HPC

Canon, Jacobsen, Henseleer

Container-based computed is rapidly changing the way software is developed, tested, and deployed. We will present a detailed overview of the design and implementation of Shifter, which in partnership with Cray has extended on the early prototype concepts and is now in production at NERSC. Shifter enables end users to execute containers using images constructed from various methods including the popular Docker-based ecosystem. We will discuss some of the improvements and implementation details. In addition, we will discuss lessons learned, performance results, and real-world use cases of Shifter in action and the potential role of containers in scientific and technical computing including how they complement the scientific process. We will conclude with a discussion about the future directions of Shifter.

08:30-10:00 Technical Session 18A

Dynamic RDMA Credentials

Shimek, Swaro

Dynamic RDMA Credentials (DRC) is a new system service to allow shared network access between different user applications. DRC allows user applications to request managed network credentials, which can be shared with other users, groups or jobs. Access to a credential is governed by the application and DRC to provide authorized and protected sharing of network access between applications. DRC extends the existing protection domain functionality provided by ALPS without exposing application data to unauthorized applications. DRC can also be used with other batch systems such as SLURM, without any loss of functionality. In this paper, we will show how DRC works, how to use DRC and demonstrate various examples of how DRC can be used. Additionally, future work for DRC will be discussed, including optimizations for performance and authorization features.

08:30-10:00 Technical Session 18B

Interactive Data Analysis using Spark on Cray Urika Appliance

Kaul, Tordoir, Petrella

In this talk, we discuss how data scientists can use the Intel Data Analytics Acceleration Library (DAAL) with Spark and Spark notebook. Intel DAAL provides building blocks for analytics optimized for x86 architecture and. We have integrated DAAL with Spark Notebook. This provides a Spark user, an interactive and optimized interface for running Spark jobs. We take real world machine learning and graph analytics workloads in bioinformatics and run it on Spark using the Cray Urika-XA appliance. The benefits of an co-designed hardware and software stack become apparent with these examples with the added benefit of an interactive front end for users.

08:30-10:00 Technical Session 18B

Experiences Running Mixed Workloads On Cray Analytics Platform

AYYALASOMAYAJULA, Maschhoff

The ability to run both HPC and big data frameworks together on the same machine is a principal design goal for future Cray analytics platforms. Hadoop provides a reasonable solution for parallel processing of batch workloads using the YARN resource manager. Spark is a general-purpose cluster-computing framework, which also provides parallel processing of batch workloads as well as in-memory data analytics capabilities; iterative, incremental algorithms; ad hoc queries; and stream processing. Spark can be run using YARN, Mesos or its own standalone resource manager. The Cray Graph Engine (CGE) supports real-time analytics on the largest and most complex graph problems. CGE is a more traditional HPC application that runs under either Slurm or PBS. Traditionally, running workloads that require different resource managers requires static partitioning of the cluster. This can lead to underutilization of resources. In this paper, we describe our experiences running mixed workloads on our next generation Cray analytics platform (internally referred to as "Athena") with dynamic resource partitioning. We discuss how we can run both HPC and big data workloads by leveraging different resource managers to interoperate with Mesos, a distributed cluster and resource manager, without having to statically partition the cluster. We also provide a sample workload to illustrate how Mesos is used to manage the multiple frameworks.

08:30-10:00 Technical Session 18B

Characterizing the Performance of Analytics Workloads on the Cray XC40

Ringenburg, Zhang, Maschhoff, Sparks, Racah, Prabhat

This paper describes an investigation of the performance characteristics of high performance data analytics (HPDA) workloads on the Cray XC40, with a focus on commonly-used open source analytics frameworks like Apache Spark. We look at two types of Spark workloads: the Spark benchmarks from the Intel HiBench 4.0 suite and a CX matrix decomposition algorithm. We study performance from both the bottom-up view (via system metrics) and the top-down view (via application log analysis), and show how these two views can help identify performance bottlenecks and system issues impacting data analytics workload performance. Based on this study, we provide recommendations for improving the performance of analytics workloads on the XC40.

08:30-10:00 Technical Session 18C

Network Performance Counter Monitoring and Analysis on the Cray XC Platform

Brandt, Froese, Gentile, Kaplan, Allan, Walsh

The instrumentation of Cray's Aries network ASIC, of which the XC platform's High Speed Network (HSN) is comprised, offers unprecedented potential for better understanding and utilization of platform HSN resources. Monitoring the amount of data generated on a large-scale system presents challenges with respect to synchronization, data management, and analysis. There are over a thousand raw counter metrics per Aries router and interface with functional combinations of these raw metrics required for insight into network state. Cray and ACES (LANL/SNL) are collaborating on collection and analysis of the Aries HSN network performance counter data. In this paper we present our work to identify HSN counters of interest; perform synchronized system wide collection of this counter data; and analyze the occurrences, levels, and longevity of congestion. We also discuss the challenges and solutions associated with collection, transport, in-transit computation and derived data analysis, visualization, storage, overhead, and redundant data.

08:30-10:00 Technical Session 18C

Dynamic Machine Specific Register Data Collection as a System Service

Bauer, Brandt, Gentile, Kot, Showerman

The typical use case for Machine Specific Register (MSR) data is to provide application profiling tools with hardware performance counter data (e.g., cache misses, flops, instructions executed). This enables the user/developer to gain understanding about relative performance/efficiencies of the code overall as well as smaller code sections. Due to the overhead of collecting data at sufficient fidelity for the required resolution, these tools are typically only run while tuning a code. In this work we present a substantially different use case for MSR data, namely system wide synchronized and relatively low fidelity collection as a system service on NCSA's 27,648 node Blue Waters platform. We present which counters we collect, the motivation for this particular data, and associated overhead. Additionally we present some associated pitfalls, how we address them, and the effects. We finally present some analysis results and the insight they provide about applications, system resources, and their interactions.

08:30-10:00 Technical Session 18C

Design and implementation of a scalable monitoring system for Trinity

DeConinck, Bonnie, Kelly, Sanchez, Martin, Mason, Brandt, Gentile, Allan, Agelastos, Davis, Berry

The Trinity XC-40 system at Los Alamos National Laboratory presents unprecedented challenges to our system management capabilities, including increased scale, new and unfamiliar subsystems, and radical changes in the system software stack. These challenges have motivated the development of a next-generation monitoring system with new capabilities for collection and analysis of system and facilities data. This paper presents the design of our new monitoring system, its implementation, and an analysis of impact on system and application performance. This will include the aggregation of diverse data feeds from the compute platform, such as system logs, hardware metrics, power and energy usage, and High Speed Network performance counters; as well as data from supporting systems such as the parallel filesystem, network infrastructure, and facilites. We will also present tools and analyses used to better understand system and application behavior, as well as insights and operational experiences from monitoring the system in production.

10:30-12:00 Technical Session 19A

What's new in Allinea's tools: from easy batch script integration and remote access to energy profiling.

Wohlschlegel

We address application energy use and performance, productivity, and the future in this talk. The Allinea Forge debugging and profiling tools, DDT and MAP, are deployed on most Cray systems - we take this opportunity to highlight important recent innovations and share the future. Creating time-efficient software whilst keeping energy consumption low calls out for tools that show performance and energy consumption simultaneously. We demonstrate how Allinea Forge enables profiling of energy usage whilst also profiling for time on Cray systems. Enabling developers and scientists to seamlessly fit tools into their lives is the theme of the second major enhancement we highlight: the new "reverse connect" capability. This removes complexities of adding debugging and profiling into complex HPC batch scripts and has been warmly received by many Cray users already. As 2016 is a busy year for new processors, we conclude by sharing upcoming plans and current status with CUG.

10:30-12:00 Technical Session 19A

Configuring and Customizing the Cray Programming Environment on CLE 6.0 Systems

Johansen

Abstract: The Cray CLE 6.0 system will provide a new installation model for the Cray Programming Environment. This paper will focus on the new processes for configuring and customizing the Cray Programming Environment to best meet the customer site's requirements. Topics will include configuring the login shell start-up scripts to load the appropriate modulefiles, creating specialized modulefiles to load specific versions of Cray Programming Environment components, and how to install third party programming tools and libraries not released by Cray. Directions will be provided on porting programming environment software to CLE 6.0 systems, including instructions on how to create modulefiles. The specific example of porting the Python MPI library (mpi4py) to CLE 6.0 will be included.

10:30-12:00 Technical Session 19A

Optimizing Cray MPI and SHMEM Software Stacks for Cray-XC Supercomputers based on Intel KNL Processors

Kandalla, Mendygral, Radcliffe, Cernohous, McMahon, Pagel

HPC applications commonly use Message Passing Interface (MPI) and SHMEM programming models to achieve high performance in a portable manner. With the advent of the Intel MIC processor technology, hybrid programming models that involve the use of MPI/SHMEM along with threading models (such as OpenMP) are gaining traction. However, most current generation MPI implementations are not poised to offer high performance communication in highly threaded environments. The latest MIC architecture, Intel Knights Landing (KNL), also offers High Bandwidth Memory - a new memory technology, along with complex NUMA topologies. This paper describes the current status of the Cray MPI and SHMEM implementations for optimizing application performance on Cray XC supercomputers that rely on KNL processors. A description of the evolution of WOMBAT (a high fidelity astrophysics code) to leverage thread-hot RMA in Cray MPICH is included. Finally, this paper also summarizes new optimizations in the Cray MPI and SHMEM implementations.

10:30-12:00 Technical Session 19B

LIOProf: Exposing Lustre File System Behavior for I/O Middleware

Xu, Byna, Venkatesan, Sisneros, Kulkarni, Chaarawi, Chadalavada

As parallel I/O subsystem in large-scale supercomputers is becoming complex due to multiple levels of software libraries, hardware layers, and various I/O patterns, detecting performance bottlenecks is a critical requirement. While there exist a few tools to characterize application I/O, robust analysis of file system behavior and associating file-system feedback with application I/O patterns are largely missing. Toward filling this void, we introduce Lustre IO Profiler, called LIOProf, for monitoring the I/O behavior and for characterizing the I/O activity statistics in the Lustre file system. In this paper, we use LIOProf for both uncovering pitfalls of MPI-IO's collective read operation over Lustre file system and identifying HDF5 overhead. Based on LIOProf characterization, we have implemented a Lustre-specific MPI-IO collective read algorithm, enabled HDF5 collective metadata operations and applied HDF5 datasets optimization. Our evaluation results on two Cray systems (Cori at NERSC and Blue Waters at NCSA) demonstrate the efficiency of our optimization efforts.

10:30-12:00 Technical Session 19B

Psync - Parallel Synchronization Of Multi-Pebibyte File Systems

Loftus

When challenged to find a way to migrate an entire file system onto new hardware while maximizing availability and ensuring exact data and metadata duplication, NCSA found that existing file copy tools couldn't fit the bill. So they set out to create a new tool. One that would scale to the limits of the file system and provide a robust interface to adjust to the dynamic needs of the cluster. The resulting tool, Psync, effectively manages many syncs running in parallel. It is dynamically scalable (can add and remove nodes on the fly) and robust (can start/stop/restart the sync). Psync has been run successfully on hundreds of nodes each with multiple processes (yielding possibly thousands of parallel processes). This talk will present the overall design of Psync and it's use as a general purpose tool for copying lots of data as quickly as possible.

10:30-12:00 Technical Session 19B

FCP: A Fast and Scalable Data Copy Tool for High Performance Parallel File Systems

Wang, Vergara Larrea, Leverman, Oral

The design of HPC file and storage systems has largely been driven by the requirements on capability, reliability, and capacity. However, the convergence of large-scale simulations with big data analytics have put the data, its usability, and management back on the front and center position. In this paper, we are introducing the FCP tool, a file system agnostic copy tool designed at the OLCF for scalable and high-performance data transfers between two file system endpoints. It provides an array of interesting features such as adaptive chunking, checksumming on the fly, checkpoint and resume capabilities to handle failures, and preserving stripe information for Lustre file system etc. It is currently available on the Titan supercomputer at the OLCF. Initial tests have shown that FCP has much better and scalable performance than traditional data copy tools and it was capable of transferring petabyte-scale datasets between two Lustre file systems.

10:30-12:00 Technical Session 19C

Performance Test of Parallel Linear Equation Solvers on Blue Waters - Cray XE6/XK7 system

Kwack, Bauer, Koric

Parallel linear equation solvers are one of the most important components determining the scalability and efficiency of many supercomputing applications. Several groups and companies are leading the development of linear system solver libraries for HPC applications. In this paper, we present an objective performance test study for the solvers available on a Cray XE6/XK7 supercomputer, named Blue Waters, at National Center for Supercomputing Applications (NCSA). A series of non-symmetric matrices are created through mesh refinements of a CFD problem. PETSc, MUMPS, SuperLU, Cray LibSci, Intel PARDISO, IBM WSMP, ACML, GSL, NVIDIA cuSOLVER and AmgX solver are employed for the performance test. CPU-compatible libraries are tested on XE6 nodes while GPU-compatible libraries are tested on XK7 nodes. We present scalability test results of each library on Blue Waters, and how far and fast the employed libraries can solve the series of matrices.

10:30-12:00 Technical Session 19C

The GNI Provider Layer for OFI libfabric

Pritchard, Harvey, Choi, Swaro, Tiffany

The Open Fabrics Interfaces (OFI) libfabric, a community-designed networking API, has gained increasing attention over the past two years as an API which promises both high performance and portability across a wide variety of network technologies. The code itself is being developed as Open Source Software with contributions from across government labs, industry and academia. In this paper, we present a libfabric provider implementation for the Cray XC system using the Generic Network Interface (GNI) library. The provider is especially targeted for highly multi-threaded applications requiring concurrent access to the Aries High Speed Network with minimal contention between threads.

10:30-12:00 Technical Session 19C

Big Data Analytics on Cray XC Series DataWarp using Hadoop, Spark and Flink

Schmidtke, Laubender, Steinke

We currently explore the Big Data analytics capabilities of the Cray XC architectures to harness the computing power for increasingly common programming paradigms for handling large volumes of data. These include MapReduce and, more recently, in-memory data processing approaches such as Apache Spark and Apache Flink. We use our Cray XC Test and Development System (TDS) with 16 diskless compute nodes and eight DataWarp nodes. We use Hadoop, Spark and Flink implementations of select benchmarks from the Intel HiBench micro benchmark suite to find suitable runtime configurations of these frameworks for the TDS hardware. Motivated by preliminary results in throughput per node in the popular Hadoop TeraSort benchmark we conduct a detailed scaling study and investigate resource utilization. Furthermore we identify scenarios where using DataWarp nodes is advantageous to using Lustre.

13:00-14:30 Technical Session 20A

Scalable Remote Memory Access Halo Exchange with Reduced Synchronization Cost

Szpindler

Remote Memory Access (RMA) is a popular technique for data exchange in the parallel processing. Message Passing Interface (MPI), ubiquitous environment for distributed memory programming, introduced improved model for RMA in the recent version of the standard. While RMA provides direct access to low-level high performance hardware, MPI one-sided communication enables various synchronization regimes including scalable group synchronization. This combination provides methods to improve performance of commonly used communication schemes in parallel computing. This work evaluates one-sided halo exchange implementation on the Cray XC40 system. Large numerical weather prediction code is studied. To address already identified overheads for RMA synchronization, recently proposed extension of Notified Access is considered. To reduce the cost of the most frequent message passing communication scheme, alternative RMA implementation is proposed. Additionally, to identify more scalable approaches, performance of general active target synchronization, Notified Access modes of RMA and original message passing implementation are compared.

13:00-14:30 Technical Session 20A

Scaling hybid coarray/MPI miniapps on Archer

Cebamanos, Shterenlikht, Arregui, Margetts

We have developed miniapps from MPI finite element library ParaFEM and Fortran 2008 coarray cellular automata library CGPACK. The miniapps represent multi-scale fracture models of polycrystalline solids. The software from which these miniapps have been derived will improve predictive modelling in the automotive, aerospace, power generation, defense and manufacturing sectors. The libraries and miniapps are distributed under BSD license, so these can be used by computer scientists and hardware vendors to test various tools including compilers and performance monitoring applications. CrayPAT tools have been used for sampling and tracing analysis of the miniapps. Two routines with all-to-all communication structures have been identified a primary candidates for optimisation. New routines have been written implementing the nearest neighbour algorithm and using coarray collectives. Scaling limit for miniapps has been increased by a factor of 3, from about 2k to over 7k cores. The miniapps uncovered several issues in CrayPAT and Cray implementation of Fortran coarrays. We are working with Cray engineers to resolve these. Hybrid coarray/MPI programming is uniquely enabled on Cray systems. This work is of particular interest to Cray developers, because it details real experiences of using hybrid Fortran coarray/MPI programming for scientific computing in an area of cutting edge research.

13:00-14:30 Technical Session 20A

Enhancing Scalability of the Gyrokinetic Code GS2 by using MPI Shared Memory for FFTs

Anton, van Wyk, Highcock, Roach, Parker

GS2 (http://sourceforge.net/projects/gyrokinetics) is a 5-D initial value parallel code used to simulate low frequency electromagnetic turbulence in magnetically confined fusion plasmas. Feasible calculations routinely capture plasma turbulence at length scales close either to the electron or the ion Larmor radius. Self-consistently capturing the interaction between turbulence at ion scale and electron scale requires a huge increase in the scale of computation. We describe a new algorithm for computing FFTs in GS2 that reduces MPI communication using MPI3 shared memory. With FFT data local to a node, the new algorithm extends perfect scaling to core-counts higher by almost a factor of 10(if the load imbalance is small). For larger FFTs we propose and analyse the performance of a version of this algorithm that distributes the FFT data in shared memory over a group of nodes with specialized MPI ranks inside each node to permit computation communication overlap.

13:00-14:30 Technical Session 20B

HPC Programming for Highly-scalable Nodes

Miles, Wolfe

High end supercomputers have increased in performance from about 4 TFLOPS to 33 PFLOPS in the past 15 years, a factor of about 10,000. Increased node count accounts for a factor of 10, and clock rate increases for another factor of 5. Most of the increase, a factor of about 200, is due to increases in single-node performance. We expect this trend to continue with single-node performance increasing faster than node count. Building scalable applications for such targets means exploiting as much intra-node parallelism as possible. We discuss coming supercomputer node designs and how to abstract the differences to enable design of portable scalable applications, and the implications for HPC programming languages and models such as OpenACC and OpenMP.

13:00-14:30 Technical Session 20B

Balancing particle and Mesh Computation in a Particle-In-Cell Code

Worley, D'Azevedo, Hager, Ku, Yoon, Chang

The XGC1 plasma microturbulence particle-in-cell simulation code has both particle-based and mesh-based computational kernels that dominate performance. Both of these are subject to load imbalances that can degrade performance and that evolve during a simulation. Each separately can be addressed adequately, but optimizing just for one can introduce significant load imbalances in the other, degrading overall performance. A technique has been developed based on Golden Section Search that minimizes wallclock time given prior information on wallclock time, and on current particle distribution and mesh cost per cell, and also adapts to evolution in load imbalance in both particle and mesh work. In problems of interest this doubled the performance on full system runs on the XK7 at the Oak Ridge Leadership Computing Facility compared to load balancing only one of the kernels.

13:00-14:30 Technical Session 20B

Computational Efficiency Of The Aerosol Scheme In The Met Office Unified Model

Richardson, O'Connor, Mann, Selwood

Abstract - A new data structuring has been implemented in the Met Office Unified Model (MetUM) which improves the performance of the aerosol subsystem. Smaller amounts of atmospheric data, in the arrangement of segments of atmospheric columns, are passed to the aerosol sub-processes. The number of columns that are in a segment can be changed at runtime and thus can be tuned to the hardware and science in operation. This revision alone has halved the time spent in some of the aerosol sections for the case under investigation. The new arrangement allows simpler implementation of OpenMP around the whole of the aerosol subsystem and is shown to give close to ideal speed up. Applying a dynamic schedule or retaining a simpler static schedule for the OpenMP parallel loop are shown to differ related to the number of threads. The percentage of the run spent in the UKCA sections has been reduced from 30% to 24% with a corresponding reduction in runtime by 11% for a single threaded run. When the reference version is using 4 threads the percentage of time spent in UKCA is higher at 40% but with the OpenMP and segmenting modifications this is now reduced to 20% with a corresponding reduction in run time of 17%. For 4 threads the parallel speed-up for the reference code was 1.78 and after the modifications it is 1.91. Both these values indicate that there is still a significant amount of the run that is serial (within an MPI task) which is continually being addressed by the software development teams involved in MetUM.

13:00-14:30 Technical Session 20C

The Hidden Cost of Large Jobs - Drain Time Analysis at Scale

Fullop

At supercomputing centers where many users submit jobs of various sizes, scheduling efficiency is the key to maximizing system utilization. With the capability of running jobs on massive numbers of nodes being the hallmark of large clusters, draining sufficient nodes in order to launch those jobs can severely impact the throughput of these systems. While these principles apply to any sized cluster, the idle node-hours due to drain on the scale of today's systems warrants attention. In this paper we provide methods of accounting for system-wide drain time as well as how to attribute drain time to a specific job. Having data like this allows for real evaluation of scheduling policies and their effect on node occupancy. This type of measurement is also necessary to allow for backfill recovery analytics and enables other types of assessments.

13:00-14:30 Technical Session 20C

The Intel® Omni-Path Architecture: Game-Changing Performance, Scalability, and Economics

Russell

The Intel® Omni-Path Architecture, Intel's next-generation fabric product line, is off to an extremely fast start since its launch in late 2015. With high-profile customer deployments being announced at a feverish pace, the performance, resiliency, scalability, and economics of Intel's innovative fabric product line are winning over customers across the HPC industry. Learn from an Intel Fabric Solution Architect how to maximize both the performance and economic benefits when deploying Intel® OPA-based cluster, and how it deliver huge benefits to HPC applications over standard Infiniband-based designs.

13:00-14:30 Technical Session 20C

How to Automate and not Manage under Rhine/Redwood

Peltz Jr., DeConinck, Grunau

Los Alamos National Laboratory and Sandia National Laboratory under the Alliance for Computing at Extreme Scale (ACES) have partnered with Cray to deliver Trinity, the Department of Energy's next supercomputer on the path to exascale. Trinity, which is an XC40, is an ambitious system for a number of reasons, one of which is the deployment of Cray's new Rhine/Redwood (CLE 6.0/SMW 8.0) system management stack. With this release came a much-needed update to the system management stack to provide scalability and a new philosophy on system management. However, this update required LANL to update its own system management philosophy, and presented a number of challenges in integrating the system into the larger computing infrastructure at Los Alamos. This paper will discuss the work the LANL team is doing to integrate Trinity, automate system management with the new Rhine/Redwood stack, and combine LANL's and Cray's new system management philosophy.

15:00-17:00 Technical Session 21A

Exploiting Thread Parallelism for Ocean Modeling on Cray XC Supercomputers

Sarje, Jacobsen, Williams, Oliker, Ringler

Incorporation of increasing core counts in modern processors used to build state-of-the-art supercomputers is driving application development towards implementation of thread parallelism, in addition to distributed memory parallelism, to deliver efficient high-performance codes. In this work we describe the implementation of threading and our experiences with it with respect to a real-world ocean modeling application code, MPAS-Ocean. We present detailed performance analysis and comparisons of various approaches and configurations for threading on the Cray XC series supercomputers, and show the benefits of threading on run time performance and energy requirements with increasing concurrency.

15:00-17:00 Technical Session 21A

Cori - A System to Support Data-Intensive Computing

Antypas, Bard, Bhimji, Declerck, He, Jacobsen, Cholia, Prabhat, Wright

The first phase of Cori, NERSC's next generation supercomputer, a Cray XC40, has been configured to specifically support data intensive computing. With increasing dataset sizes coming from experimental and observational facilities, including telescopes, sensors, detectors, microscopes, sequencers, and, supercomputers, scientific users from the Department of Energy, Office of Science are increasingly relying on NERSC for extreme scale data analytics. This paper will discuss the Cori Phase 1 architecture, and installation into the new and energy efficient CRT facility, and explains how the system will be combined with the larger Cori Phase 2 system based on the Intel Knights Landing processor. In addition, the paper will describe the unique features and configuration of the Cori system that allow it to support data-intensive science.

15:00-17:00 Technical Session 21A

Stitching Threads into the Unified Model

Glover, Selwood, Malcolm, Guidolin

The Met Office Unified Model (UM) uses a hybrid parallelization strategy: MPI and OpenMP. Being legacy code, OpenMP has been retrofitted in a piecemeal fashion over recent years. As OpenMP coverage expands, we are able to perform operational runs on fewer MPI tasks for a given machine resource, achieving the aim of reduced communication overheads. Operationally, we are running with 2 threads today; but we are on the cusp of 4 threads becoming more efficient for some model configurations. Outside of operational resource windows, we have been able to scale the UM to 88920 cores, which would not have been possible with MPI alone. We have experimented with a resource-stealing strategy to mitigate load imbalance between co-located MPI tasks. The method is working from a technical point of view and showing large performance benefits in the relevant code, but there are significant overheads which require attention.

15:00-17:00 Technical Session 21A

On Enhancing 3D-FFT Performance in VASP

Wende, Marsman, Steinke

We optimize the computation of 3D-FFT in VASP in order to prepare the code for an efficient execution on multi- and many-core CPUs like Intel's Xeon Phi. Along with the transition from MPI to MPI+OpenMP, library calls need to adapt to threaded versions. One of the most time consuming components in VASP is 3D-FFT. Beside assessing the performance of multi-threaded calls to FFTW and Intel MKL, we investigate strategies to improve the performance of FFT in a general sense. We incorporate our insights and strategies for FFT computation into a library which encapsulates FFTW and Intel MKL specifics and implements the following features: 1) reuse of FFT plans, 2) composed FFTs, 3) high bandwidth memory on Intel's KNL Xeon Phi, 4) auto-tuning based on runtime statistics. We will present results on a Cray-XC40 and a Cray-XC30 Xeon Phi system using synthetic benchmarks and with the library integrated into VASP.

15:00-17:00 Technical Session 21B

The time is now. Unleash your CPU cores with Intel® SSDs

Kudryavtsev, Furnanz Andrey Kudryavtsev

HPC Solution Architect for the Intel® Non-Volatile Solutions Group (NSG) will discuss advancements in Intel SSD technology that is unleashing the power of the CPU. He will dive into the benefits of Intel® NVMe SSDs that can greatly benefit HPC specific performance with parallel file systems. He will also share the HPC solutions and performance benefits that Intel has already seen with their customers today, and how adoption of the current SSD technology sets the foundation for consumption of Intel's next generation of memory technology 3D Xpoint Intel® SSDs with Intel Optane™ Technology in the High Performance Compute segment.

15:00-17:00 Technical Session 21B

Introducing a new IO tier for HPC Storage

Coomer

Tier 1 "performance" storage is becomingly increasingly flanked by new, solid-state-based tiers and active archive tiers that improve the economics of both performance and capacity. The available implementations of solid-state tiers into parallel filesystems are typically based on a separate namespace and/or utilise existing filesystem technologies. Given the price/performance characteristics of SSDs today, huge value is gained by addressing both optimal data placement to the SSD tier and in comprehensively building this tier to accelerate the broadest spectrum of IO, rather than just small IO random read. Dr. James Coomer, Technical Director for DDN Europe, will present the background collaboration, key technologies and community work - that has led to the development of an entirely new class of cache with extreme scalability that maximises the utility of NVMe at scale in HPC and optimises the IO to parallel file systems and archive tiers. The talk will include several case studies.

15:00-17:00 Technical Session 21B

H5Spark: Bridging the I/O Gap between Spark and Scientific Data Formats on HPC Systems

Liu, Racah, Koziol, Canon, Gittens, Gerhardt, Byna, Ringenburg, Prabhat

Spark has been tremendously powerful in performing Big Data analytics in distributed data centers. However, using the Spark framework on HPC systems to analyze large-scale scientific data has several challenges. For instance, parallel file system is shared among all computing nodes in contrast to shared-nothing architectures. Another challenge is in accessing data stored in scientific data formats, such as HDF5 and NetCDF, that are not natively supported in Spark. Our study focuses on improving I/O performance of Spark on HPC systems for reading and writing scientific data arrays, e.g., HDF5/netCDF. We select several scientific use cases to drive the design of an efficient parallel I/O API for Spark on HPC systems, called H5Spark. We optimize the I/O performance, taking into account Lustre file system striping. We evaluate the performance of H5Spark on Cori, a Cray XC40 system, located at NERSC.

15:00-17:00 Technical Session 21C

Maintaining Large Software Stacks in a Cray Ecosystem with Gentoo Portage

MacLean

Building and maintaining a large collection of software packages from source is difficult without powerful package management tools. This task is made more difficult in an environment where many libraries do not reside in standard paths and where loadable modules can drastically alter the build environment, such as on a Cray system. The need to maintain multiple Python interpreters with a large collection of Python modules is one such case of having a large and complicated software stack and is described in this paper. To address limitations of current tools, Gentoo Prefix was ported to the Cray XE/XK system, Blue Waters, giving the ability to use the Portage package manager. This infrastructure allows for fine-grained dependency tracking, consistent build environments, multiple Python implementations, and customizable builds. This infrastructure is used to build and maintain over 400 packages for Python support on Blue Waters for use by its partners.

15:00-17:00 Technical Session 21C

Evaluating Shifter for HPC Applications

Bahls

Shifter is a powerful tool that has the potential to expand the availability of HPC applications on Cray XC systems by allowing Docker-based containers to be run with little porting effort. In this paper, we explore the use of Shifter as a means of running HPC applications built for commodity Linux clusters environments on a Cray XC under the Shifter environment. We compare developer productivity, application performance, and application scaling of stock applications compiled for commodity Linux clusters with both Cray XC tuned Docker images as well as natively compiled applications not using the Shifter environment. We also discuss pitfalls and issues associated with running non-SLES-based Docker images in the Cray XC environment.

15:00-17:00 Technical Session 21C

Executing dynamic heterogeneous workloads on Blue Waters with RADICAL-Pilot

Santcroos, Castain, Merzky, Bethune, Jha

Traditionally HPC systems such as Cray's have been designed to support mostly monolithic workloads. The workload of many important scientific applications however, is constructed out of spatially and temporally heterogeneous tasks that are often dynamically inter-related. These workloads can benefit from being executed at scale on HPC resources but a tension exists between the workloads resource utilization requirements and the capabilities of the HPC system software and usage policies. Pilot systems have the potential to relieve this tension. In this paper we describe RADICAL-Pilot, a scalable and interoperable pilot system that enables the execution of these diverse workloads. We describe the design and characterize the performance of its task executing components, which are engineered for efficient resource utilization while maintaining the full generality of the Pilot abstraction. We will discuss three different implementations of support for RADICAL-Pilot on Cray systems and analyse their performance.

15:00-17:00 Technical Session 21C

Early Application Experiences on Trinity - the Next Generation of Supercomputing for the NNSA

Vaughan, Dinge, Lin, Pierson, Hammond, Cook, Trott, Agelastos, Pase, Benner, Rajan, Hoekstra

Trinity, a Cray XC40 supercomputer, will be the flagship capability computing platform for the United States nuclear weapons stockpile stewardship program when the machine enters full production during 2016. In the first phase of the machine, almost 10,000 dual socket Haswell processor nodes will be deployed, followed by a second phase utilizing Intel's next-generation Knights Landing processor. In this talk we will describe production scientific and engineering application performance baselining on test machines for the Phase I compute partition comparing our data to the previous XE6 capability machine, Cielo. The talk will contain discussion of our first experiences with the machine and performance results which showcase significantly improved compute performance and application MPI behavior including message sizes and bandwidths.

17:05-17:35 General Session 22

Conference Closing Session

Hancock

Closing Remarks Thank-You's CUG 2017 Announcement

Social Events

Monday, 9th May 2016

18:30 – 21:30 Dickens Inn Special Event Sponsored by TBC

Chair: TBC

www.dickensinn.co.uk

Walk west on Prescot St/B126 towards W Tenter St. Slight left onto Mansell St/A1210, continue onto Tower Bridge Approach/A100, slight left onto St Katharine's Way, turn left and take the stairs towards Mews St. Follow the dock around to the left, then right and Dickens Inn will be situated on your left. The private event will be held in the left bar, Copperfield Bar. Drinks will be served upon arrival with food at approximately 19:30.

Tuesday, 10th May 2016

18:30 – 21:30 Cray Networking Event at The Folly

Chair: Christy Adkinson (Cray, Inc)

http://www.thefollybar.co.uk/

Attendees should walk west on Prescot St/B126 towards W Tenter St, take a slight right onto Mansell St/A1210. Turn left onto Portsoken St and continue onto Crosswall. Turn right onto Crutched Friars, turn left onto Lloyd's Ave, turn left onto Fenchurch Street then turn left onto Gracechurch St/A1213. Take a slight right to stay on Gracechurch St/A1213 where The Folly will be found on your right. Alternatively, attendees can take the District Line from Tower Hill station to Monument where The Folly is approximately 1 minute walk.

Wednesday, 11th May 2016

19:00 – 23:00 CUG Night out at The Museum

Chair: Jim Rogers (Oak Ridge National Laboratory)

http://www.museumoflondon.org.uk/london-wall/

CUG Night Out at the Museum of London. Transportation from The Grange to the museum and back will be provided. Buses will load at The Grange between 18:45 and 19:00. The event starts at 7:00 PM at the museum. Coat and bag check will be available upon arrival. A variety of reception canapes will be served in the Galleries of the museum, allowing you to experience London's history from 450,000 BC right up to the present day. After you have taken in the amazing story of London and its people, you will take your seats for a three course meal in Sackler Hall. Shuttle transport has been arranged for the return journey. Buses will depart at 22:00, 22:20, 22:40 and 23:00.

Local Arrangements

How to Contact Us

After the conference:

Oak Ridge National Laboratory
Attn: Jim Rogers
1 Bethel Valley Road P.O. Box 2008; MS 6008
Oak Ridge, TN 37831-6008
cug2016@cug.org

During the conference:

You can find us in The Prescott Suite (aka CUG Office) on the 1st floor or at the registration desk, also on the 1st floor.

Conference Registration

Jim Rogers
Oak Ridge National Laboratory
1 Bethel Valley Road Oak Ridge, TN 37831-6008 USA
(1-865)-576-2978 Fax: (1-865)-241-9578
jrogers@ornl.gov

Attendance and Registration

Badges and registration materials will be available:

Sunday: 3:00 p.m. to 6:00 p.m.
Monday: 7:30 a.m. to 5:30 p.m.
Tuesday: 7:30 a.m. to 10:30 a.m.
Registration desk – 1st floor

To register after Tuesday morning visit the CUG office on the 1st floor.

All attendees must wear badges during CUG Conference activities.

Smoking Policy

There is no smoking allowed at the Conference.

Special Assistance

Any requests for special assistance during the conference should be noted on the “Special Requirements” area of the registration form

Conference Registration Fees

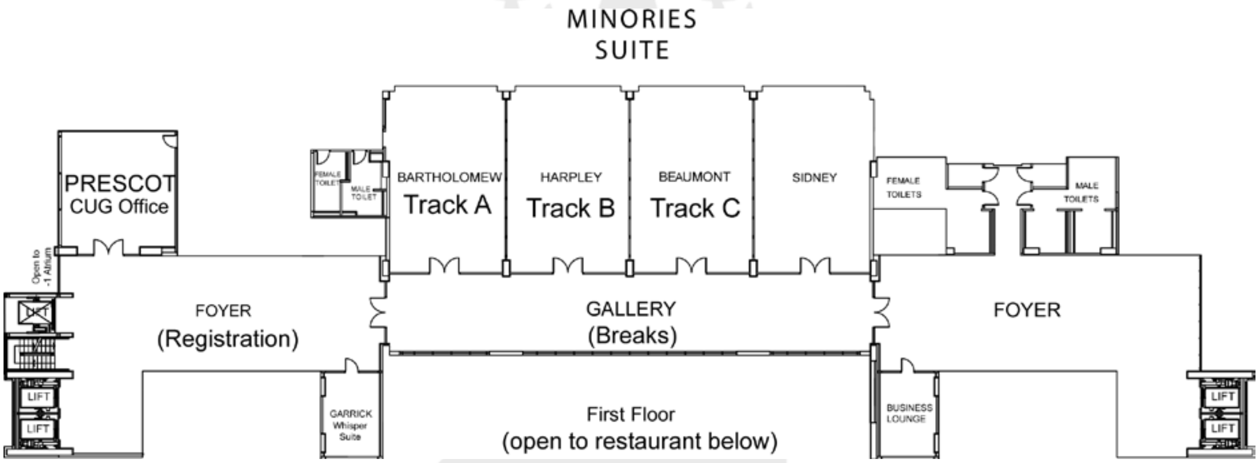
Your registration fee includes

- Admission to all program sessions, meetings, and tutorials
- Arrival, Morning and afternoon breaks, and lunch Monday through Thursday
- CUG Night Out on Tuesday night

Proceedings

Proceedings details will be announced at the conference. Sites can use their member login or contact board@cug.org for general access

The Grange Floor Plan



Sponsors

2016 HOST

ECMWF
EUROPEAN CENTRE FOR MEDIUM RANGE WEATHER FORECASTS

Diamond Sponsor

CRAY®
THE SUPERCOMPUTER COMPANY

Platinum Sponsor

intel®

Special Event Sponsor — Welcome Reception Sponsor — Gold Sponsor

DDN® STORAGE

SEAGATE

NVIDIA.

Silver Sponsors

Adaptive COMPUTING

allinea

Altair

SchedMD

SPECTRA

UNIVA

Bronze Sponsor

Bright Computing

Contacts

CUG Board

President
David Hancock
Indiana University

Vice-President
Andrew Winfer
King Abdullah University of Science & Technology

Secretary
Tina Declerck
National Energy Research Scientific Computing Center

Treasurer
Jim Rogers
Oak Ridge National Laboratory

Director-at-Large
Richard Barrett
Sandia National Laboratory

Director-at-Large
Liam Forbes
Arctic Region Supercomputing Center

Director-at-Large
Open

Past President++
Nicholas Cardo
Swiss National Supercomputer Centre

Cray Advisor to the CUG Board **
Christy Adkinson
Cray Inc.

** Note: This is not a CUG Board position. ++ Note: Appointed Position

EMAIL board@cug.org for general CUG inquiries or cug2016@cug.orgfor specific inquiries.

Special Interest Groups

Programming Environments, Applications and Documentation

Chair: Tim Robinson (CSCS)
Deputy Chair: Ashley Barker (ORNL)
Deputy Chair: Helen He (NERSC)
Deputy Chair: Rolf Rabenseifner (HLRS)
Deputy Chair: Greg Bauer (NCSA)
Deputy Chair: Suzanne Parete-Koon (ORNL)
Deputy Chair: Zhengji Zhao (NERSC)
Cray Inc. SIG Liaison Applications: Jef Dawson
Programming Environments: Luiz DeRose
Documentation: Kevin Stellj

Systems Support

Chair: Jason Hill (ORNL)
Deputy Chair: Hans-Hermann Frese
Deputy Chair: Sharif Islam
Cray Inc. SIG Liaison Systems & Integration, Operating Systems, and Operations: Janet Lebens

XTreme Systems

Chair: Tina Butler (NERSC)
Deputy Chair: Frank Indiviglio (NCRC)
Cray Inc. SIG Liaison: Chris Lindahl

CUG 2017

2017
CUG.2017.CAFFEINATED COMPUTING
Redmond, Washington May 7-11, 2017

CRAY USER GROUP 2017 I Redmond, WA, USA

The Cray User Group board would like to invite you to Redmond Washington for CUG 2017 hosted by Cray and the CUG Board May 7-11, 2017.

The 60th CUG has us returning to the Seattle area near Cray headquarters and in the heart of Washington wine-tasting country. As we look forward to 2017 the goal of the CUG Board is to expand the horizons of our attendees and our program to include even more data, storage, and analytics discussions while continuing to support high quality HPC content that is at the core of CUG.

There are many challenges in our field from rising power utilization, increasing core counts and scalability limitations, to the integration of new approaches that couple data analytics, HPC, and containerization in single environments. It is through events like CUG that we can share our approaches, expertise, and failures in order to continue to push boundaries. The relatively young history of Seattle has striking parallels to our own industry, in how difficult it was to first settle and the success it has achieved in the Pacific Northwest. Home to leading industry such as Boeing, Microsoft, and Amazon there's much we can learn from the area. Most importantly, to many sleep-deprived tech workers, it also has a thriving coffee culture, including being the birthplace of Starbucks. The theme of Caffeinated Computing touches on that tradition, and it may not be a coincidence that Seattle has such a storied history with coffee along with being the headquarters to our dear colleagues at Cray.

We value your contributions to our technical program and look forward to the progress that members and sponsors will share at CUG 2017.

David Hancock
President, Cray User Group